Effective crowd anomaly detection through spatio-temporal texture analysis

HAO, Yu, XU, Zhi-Jie, LIU, Ying, WANG, Jing and FAN, Jiu-Lun

International Journal of Automation and Computing

DOI: 10.1007/s11633-018-1141-z

Effective Crowd Anomaly Detection Through Spatio-temporal Texture Analysis

Yu Hao[1,2] Zhi-Jie Xu[2] Ying Liu[1] Jing Wang[3] Jiu-Lun Fan[1]

[1] School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China

[2] School of Computing and Engineering, University of Huddersfield, Huddersfield HD1 3DH, UK

[3] Faculty of Arts Computing Engineering and Sciences, Sheffield Hallam University, Sheffield S1 1WB, UK

Abstract: Abnormal crowd behaviors in high density situations can pose great danger to public safety. Despite the extensive installation of closed-circuit television (CCTV) cameras, it is still difficult to achieve real-time alerts and automated responses from current systems. Two major breakthroughs have been reported in this research. Firstly, a spatial-temporal texture extraction algorithm is developed. This algorithm is able to effectively extract video textures with abundant crowd motion details. It is through adopting Gabor-filtered textures with the highest information entropy values. Secondly, a novel scheme for defining crowd motion patterns (signatures) is devised to identify abnormal behaviors in the crowd by employing an enhanced gray level co-occurrence matrix model. In the experiments, various classic classifiers are utilized to benchmark the performance of the proposed method. The results obtained exhibit detection and accuracy rates which are, overall, superior to other techniques.

Keywords: Crowd behavior, spatial-temporal texture, gray level co-occurrence matrix, information entropy.

1 Introduction

Closed-circuit television (CCTV) cameras are widely installed in city centers, along main roads and highways, fixed and/or moving locations inside stadiums, concert halls, shopping malls, and other key installations for ensuring public welfare and safety. The live video feeds are often sent to various control centers for processing and storage. If the monitored crowds exhibit unusual behavioral (motion) patterns, immediate actions can be taken in response, to avoid potential damage or even casualties. For example, when the population density of a crowd in a public event is rapidly increasing and reaching a threshold, measures might need to be taken quickly to avoid a stampede; or, when people in a tightly packed tube station suddenly disperse and run away, an alarm needs to be immediately triggered in the control room. However, the main operational mode today in many countries still relies on human operators to constantly monitor live video streams from multiple sources. This is often in the form of a multi-screen monitor wall, which is a tedious job that easily leads to fatigue, slow-response or even oversight, not to mention the cost of staffing. The primary goal of this research is to design an automatic detection system which could alert human operators to the occurrence of abnormal crowd events, or even predict them.

Many approaches have been proposed for designing crowd behavioral analysis algorithms over the last two decades[1–7]. The main objectives of analyzing crowd behaviors focus on two topics: global scale (or macroscopic) analysis, local scale (or microscopic) analysis. In global scale analysis, the crowd of similar motions is treated as a single entity. Its main goal is to recognize the dominant and/or anti-dominant patterns of this entity, without concerning itself with any individual behaviors. For example, the congestion or stampede scenarios are a convergence of a crowd's locomotion. The global scale analysis, therefore, concentrates on the overall tendencies of the critical mass rather than specific behavior such as waving or jumping. In local scale analysis, the detection of an individual behavior, or more specifically, actions, among other crowd entities becomes a focus, and poses a challenging question, especially when crowd density is high. This includes, e.g., occlusions that make the segmentation of a particular individual a challenging task.

For global feature-based approaches, feature patterns such as optical flow are often extracted from entire video footage, and corresponding histograms are constructed. In the bag of visual word (BoW) technique[8], histograms with similar patterns are clustered to train a dictionary, and then the crowd behavior in a testing video is classified with its histogram. Solmaz et al.[1] proposed an algorithm to identify crowd behaviors based on optical flow information. In their research, the optical flow method is

Research Article

Special Issue on Addressing Global Changlleges through Automation and Computing

Manuscript received March 13, 2018; accepted June 6, 2018

Recommended by Associate Editor Jie Zhang

reproduced and evaluated, and then optimization work is carried out to introduce the particle angles as a new parameter for sorting and clustering the so-called regions of interests (RoI) model. By investigating the signature values calculated from the Jacobian matrix of pixel values in each RoI, different behavioral types can then be determined. Krausz and Bauckhage[2] followed a different route in tackling the problem by computing the histograms of the motion direction and magnitude extracted from the optical flow through applying the non-negative matrix factorization (NMF). The obtained histograms are then readily clustered. The essence of the process relies on a signature named as the symmetry value being calculated on the averaged histograms to check if the current cluster is in a congested state or otherwise. For local-feature-based approaches, each individual is treated as a single agent and its motion analyzed independently. One typical approach is the social force model (SFM) proposed by Helbing and Molnar[9]. The assumption of SFM is that the behaviors of each agent in a crowd are determined by multiple types of interaction forces. The extracted flow-based feature is mapped to each agent according to the rules of SFM to define individuals′ abnormal behaviors. Yan et al.[10] proposed a technique using SFM to detect sudden changes in crowd behavior. In this approach, the interaction force in SFM is directly calculated from the code stream to increase efficiency, then the BoW algorithm is applied to generate histograms on intensity and angles of interaction force flow. With the histograms obtained, the crowd′s moving state can be distinguished to detect the anomalies.

Despite the varied approaches mentioned above, the common pitfall of them is the heavy time consumption of calculating optical flow for every frame[11]. In order to maintain the detection accuracy while keeping the workload as low as possible, spatio-temporal information is explored in this research, with the aim of developing a practical crowd anomaly detection and classification framework.

Spatio-temporal information is widely used for single human action recognition, such as gesture, gait, and pose estimation. Niyogi and Adelson[12] used spatio-temporal texture (STT) to analyze human walking patterns, such as gaits at the ankle level. In this research, the key patterns of gaits were firstly defined as various braided streaks extracted from STT, and then the rough estimation of the walker′s pattern was refined using snakes (modeled streaks) proposed by Kass et al.[13]. The walker′s body was modeled by merging the Snake contours into one before the general combinatory contour was classified using the predefined gait signatures. In Wang′s research[14], dynamic events and actions were modeled and represented by various geometrical and topological structures extracted from identified spatio-temporal volumes (STV) in a scene. Similar to the individual′s behavior, crowd behavior would also generate abundant motion patterns in the spatio-temporal space. Hence, by extracting the spatio-temporal information from regions-of-interest (RoIs) in a crowd, background and irrelevant information can be culled thus saving precious computational time. In recent research by Van Gemeren[15], a novel model is proposed to detect the interaction of two persons in unsegmented videos using spatio-temporal localization. In this research, the spatio-temporal information is utilized to help model the person′s body pose and motion in detailed coordination with designed part detectors. The researcher claims to have obtained robust detection results when training on only small numbers of behavioral sequences. Ji et al.[16] introduced an approach using the combination of local spatio-temporal features and global positional distribution information to extract 3-dimensional (3D) scale-invariant feature transform (SIFT) descriptors on detected points-of-interest. Then, the SVM is applied to the descriptor for human action classification and recognition.

An abstract pipeline of the crowd anomaly detection framework proposed in this research is shown in Fig. 1. Once the raw video data is acquired, the first phase of the procedure is to perform the preprocessing operations, including noise filtering and background subtraction. Initial steps for the construction of STVs from raw video data also occur at this stage. In the second phase, main crowd features and patterns are extracted from the filtered data, where the features are modeled as descriptors (or signature vectors) for the classification/recognition purpose. In the third phase, extracted crowd patterns are sorted using various machine-learning models such as classifiers and templates. Once the crowd behaviors are identified, the abnormal ones can be treated as anomalies in further studies such as semantic analysis.

This paper is organized as follows: Section 2 introduces a novel model for identifying and extracting spatial-temporal textures (STT) from video footage.

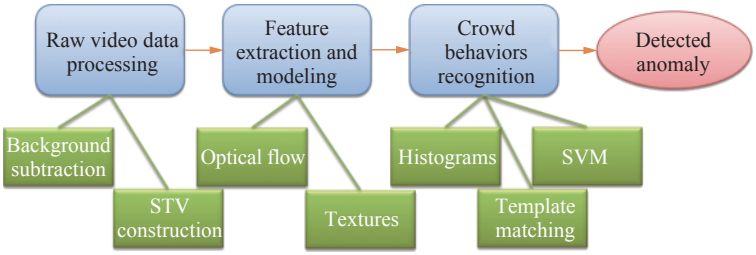


Fig. 1 A general structure of crowd abnormal behavior detection system

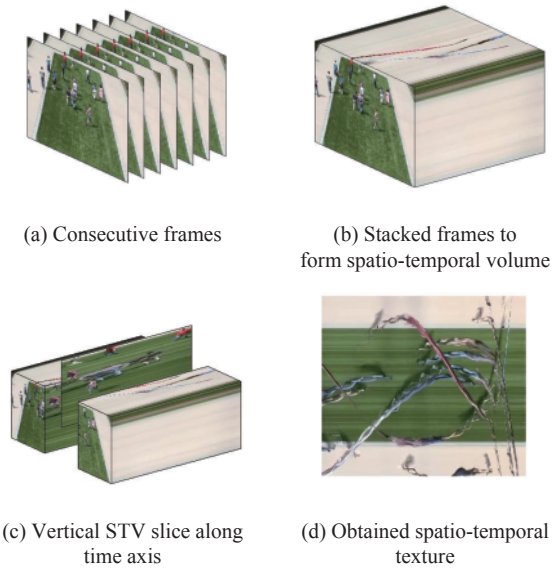
Section 3 defines a salient STT signature using a gray level co-occurrence matrix for crowd anomaly labeling. Section 4 presents the experimental results of using the proposed gray level co-occurrence matrix (GLCM) signature on various classifiers. Section 5 concludes the paper.

2 Effective spatio-temporal texture extraction

Because automatic classification of crowd patterns includes abrupt and abnormal changes, a novel approach for extracting motion "textures" from dynamic STV blocks formulated by live video streams has been proposed. This section starts by introducing the common approach for STT construction and corresponding spatio-temporal texture extraction techniques. Next, the crowd motion information contained within the random STT slices is evaluated based on the information entropy theory to cull the static background and noises occupying most of the STV spaces. A preprocessing step using Gabor filtering for improving the STT sampling efficiency and motion fidelity has been devised and tested. The technique has been applied on benchmarking video databases for proof-of-concept and performance evaluation. Preliminary results have shown encouraging outcomes and promising potential for its real-world crowd monitoring and control applications, detailed in Section 4.

2.1 STV-based motion encapsulation and STT feature representation

STV is first proposed by Aldelson and Bergen[17]. Fig. 2 illustrates the STV construction process. The live video signal is first digitized and stored as continuous and evolving 3-dimensional (3D) STV blocks. The construction of a typical STV block from video can be described as the stacking up of consecutive video frames to a fixed time capsule (normally of a few seconds) that consists of evenly spread grey-scale (for black-and-white video) or colored (for color video) mini-cubes over the 3D space, enclosed by the borders of the frame and the length (decided by the STV length in seconds and the video frame rate) along the time axis (Fig. 2(a)). Actually those cubes are 2D pixels of each frame "stretched" into 3D voxels (volumetric-pixels) filling up the STV block (Fig. 2(b)). Compared to 2D frames, a STV block naturally encapsulates dynamic information, such as object movements, as well as static scene information in its structure. 2D neighboring frame-based tracking techniques such as the optical flow[18] study the consecutive frame pairs for gradual object motions that work well for continuous human and vehicle tracking. However, this technique has major drawbacks when it comes to evaluating sudden changes, especially concerning a large group of fast moving objects within a dense crowd. In order to further process the constructed STVs, slices of a STV called spatio-temporal textures (STTs) can be extracted to learn patterns recorded



(a) Consecutive frames (b) Stacked frames to form spatio-temporal volume

(c) Vertical STV slice along time axis (d) Obtained spatio-temporal texture

Fig. 2 Procedures to obtain STV and STT from raw video data

in each texture, resembling the medical operations of 3D ultrasonic scan or magnetic resonance imaging (MRI). For example, Niyogi and Adelson[12] used STTs to analyze the gait (walking style) of individual pedestrian. In Fig. 2(c), STV is sliced either horizontally or vertically at certain position along time axis, to obtain STTs, and Fig. 2(d) shows an example of extracted STTs describing pedestrians′ motion through time.

STV and STT techniques have been widely studied in the last two decades. Bolles et al.[19] used STV for geometric and structure recovery from static scenes. Baker et al.[20, 21] used STV for 3D scene segmentation. Ngo et al.[22] used STT techniques for the detection of camera cuts, wipes and dissolves in a video sequence. In this approach, a STT was analyzed by first convolving with the first derivative Gaussian, and then processed using Gabor decomposition, in which the real components of multiple spatial-frequency channel envelopes were retrieved to form the texture feature vector. A Markov energy-based image segmentation algorithm was then used to locate the color and texture discontinuities at region boundaries. The approach was tested on different types of videos, including news and movies. The results show sound performance on "cut" detection with accuracy reaching 95%, but only 64% for the "wipe" detection.

Because of the way a STV block is constructed and the random nature of real-life events, the "useful" information distributed over a STV space is usually uneven and irregular. Thus, one important problem is how to obtain the STT slices from a STV block with the highest information density. Core to the challenge is how to differentiate useful information such as voxels formed by crowd movement from noise such as static background. In this research, instead of an even cut and computation on all STT slices from a STV block, an optimized technique is developed to obtain the specific STT with rich motion information as shown in Fig. 3.

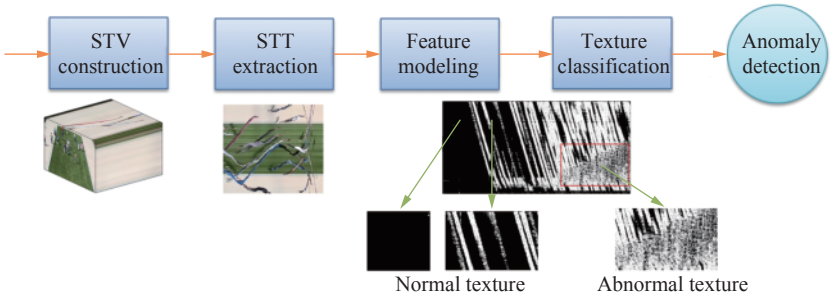


Fig. 3 Optimized framework of crowd behavior detection based on spatio-temporal information

2.2 Implementation strategy

A typical pipeline of the crowd abnormality detecting system contains three processing phases[23] as shown in Fig. 1. In the first video data acquisition phase, the raw video signals are collected and stored in suitable digital formats. Then, static or dynamic features contained within the information packets will be extracted; and at last, predefined feature patterns describing signal-level, statistical-level, and/or even semantic-level explanations of the "video events" will be used to evaluate the similarity and differences of the features extracted from the live feeds[24–26].

In this research, at the STT extraction phase, an information entropy evaluation model has been devised to help the sampling and selection of "meaningful" feature containers before feeding them into the feature (crowd patterns) extraction module. This design ensures the STT that contains the most of the crowd dynamics will be selected based on the magnitude and richness of motion "trails" along the time axis in the continuously evolving STV blocks. After that, motion features are extracted from the selected STTs and are modeled into feature vectors (signatures). In the last step of the devised framework, the identified STT RoIs are classified according to their motion signatures.

2.3 Information entropy-based STT selection

Information entropy (also referred as Shannon entropy) is proposed by Shannon[27]. It is a concept from information theory that calculates how much information there is in an event. The information gain is a measure of the probability of a certain result to occur[28]. Liang et al.[29] proposed an approach to detect encoded malicious web pages based on their information entropy counts. Zhang et al.[30] used information entropy to detect mobile payment anomaly through recursively training devised entropy mechanism using verified data. The idea of information entropy could also be used as an index to measure the informational value of the extracted STTs. If a STT has higher entropy, it is likely to contain higher motion and scene update information.

As illustrated in Fig. 2, multiple horizontal and vertical cuts can be applied to a STV block for obtaining STTs. All of the cuts are along the time axis. The sampling density of the cuts is customizable and depends on actual application scenarios. When the density is set to a higher value, it can be predicted that the result would be closer to optimal, yet the computational burden will increase. In the third step of Fig. 2, once the STTs are obtained, the information entropy is calculated for each STT. The slice with the highest information entropy will then be selected as the target STT for crowd behavior analysis.

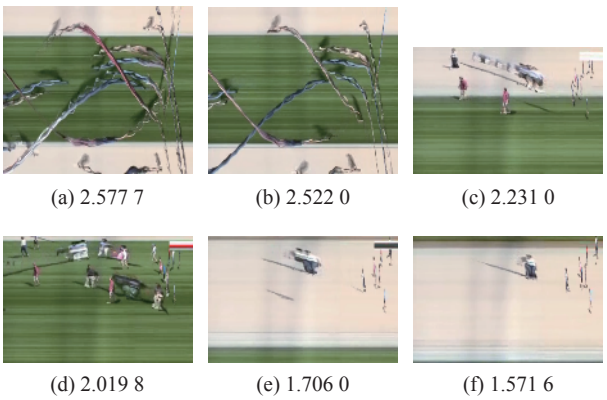
The information entropy can be expressed as (1).

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i). \tag{1}$$

In (1), n represents the total number of different gray scale levels in a STT, x_i represents the amount of pixels of the gray scale level i in it, $P(x_i)$ represents the probability of gray scale level i in the STT, and $H(X)$ is the calculated information entropy.

Fig. 4 shows the calculated information entropy values of a group of extracted STTs from a single STV. The STTs are displayed in descending order according to the calculated entropies. It can be observed that STTs with higher information entropy show abundant motion information as indicated by the ribbon-shape trajectories.

However, when directly applied to a test video database as shown in Table 1, the immediate results do not

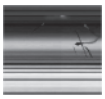
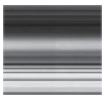
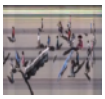


(a) 2.577 7 (b) 2.522 0 (c) 2.231 0

(d) 2.019 8 (e) 1.706 0 (f) 1.571 6

Fig. 4 Entropy values of random STTs

Table 1 Results of selected target STTs' information entropy values

	UMN1	UMN2	UMN3	UMN4	UMN5	UMN6	UMN7
STT							
Entropy	2.5777	2.6929	3.2176	3.2494	3.2404	2.6951	2.7248

seem to yielding consistent and satisfactory outcomes against intuition, where UMN3, UMN5 and UMN6 even show higher entropy values yet contain less motion features than UMN1 and UMN2.

2.4 Optimization through Gabor filtering

In Section 2.3, the information entropy is calculated on all extracted STTs, the STT with largest entropy would be selected as target for further pattern analysis. However, preliminary tests have shown unsatisfactory pairing between STT slices with high entropy values from the ones actually containing more crowd motion "ribbons". Close inspection revealed that the main cause of the problem is due to the traces left on STTs caused by non-moving objects and background regions, especially those with high color contrast. For example, the obtained sample STTs from UMN3 to UMN8 patches have shown explicit parallel stripes caused by the background. To address this issue, in this research, the Gabor wavelet filtering is exploited for removing the STT background. Fig. 5 shows the renovated processes. Instead of applying the information entropy calculation directly on the extracted STTs, they are firstly converted into gray scale images. Then, the background of STTs is removed through implementing the convolutions of the STTs with the Gabor filter before the entropy measures are calculated.

The Gabor transformation is a special case of the short-time Fourier transformation. Because the Gabor wavelet is very similar to a single cell's response to visual stimulus from the human vision system, it is sensitive to the border of an image, but not so much so to the change of light, which made it ideal in many application areas in image processing and computer vision. Panda and Meher[31] introduced a hierarchical algorithm for both block-based and pixel-based background subtraction approaches based on the Gabor transformed magnitude feature. Zhou et al.[32] extracted features using circular Gabor filters at five different frequencies, to solve the challenge that conventional background subtraction algorithms struggle to achieve.

In the spatial domain, a two dimensional Gabor filter is the product of a sinusoidal function and a Gaussian function, it is also called the window function. In practice, the Gabor filter can extract features from multiple scales and orientations. For this research, it is expressed as

$$G(x,y,\theta,f) = \exp\left(-\frac{1}{2}\left(\left(\frac{x'}{sx}\right)^2+\left(\frac{y'}{sy}\right)^2\right)\right)\times\cos(2\pi f x'). \quad (2)$$

In (2), sx and sy are the window sizes along x and y axis, and the value of x varies from negative sx to positive sx, the value of y varies from negative sy to positive sy. θ defines the orientation of the extraction process. f defines the frequency of the sinusoidal function. And

$$\begin{aligned} x' &= x\cos\theta + y\sin\theta \\ y' &= y\cos\theta + x\sin\theta. \end{aligned} \quad (3)$$

The convolution of the Gabor filter and an original STT is then applied to obtain the filtered version.

In a real-life scenario, the motion of crowd recorded in a STV block could be towards any direction, thus the Gabor filtering is applied in eight directions (like the notions of N, S, E, W, NE, SE, NW and SW on a map) to increase the accuracy. Fig. 6 shows the detailed steps of the procedure. The first and second row illustrate the filtered STTs in eight orientations respectively. Note that the parameters of Gabor filter are adjusted accordingly. In this case, values of sx and sy are set to 2, and f is set to $\sqrt{4.99}$ on Fig. 6(b)–6(e) and $\sqrt{3.9}$ on Fig. 6(g)–6(j). Once the filtering steps are completed, all 8 filtered STTs are accumulated together to formulate a combined one as shown in Fig. 6(f), where Fig. 6(a) is the original STT.

By using this method, the long computational time of calculating flow-based information in every frame can be greatly shortened. The extraction of flow-based information involves the calculation on every pixel in the video data. The amount of pixels needing to be analyzed is wht, therefore the computational complexity is $O(n^3)$.

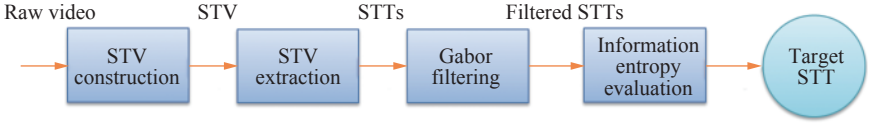


Fig. 5 Updated structure of the proposed STT extraction technique

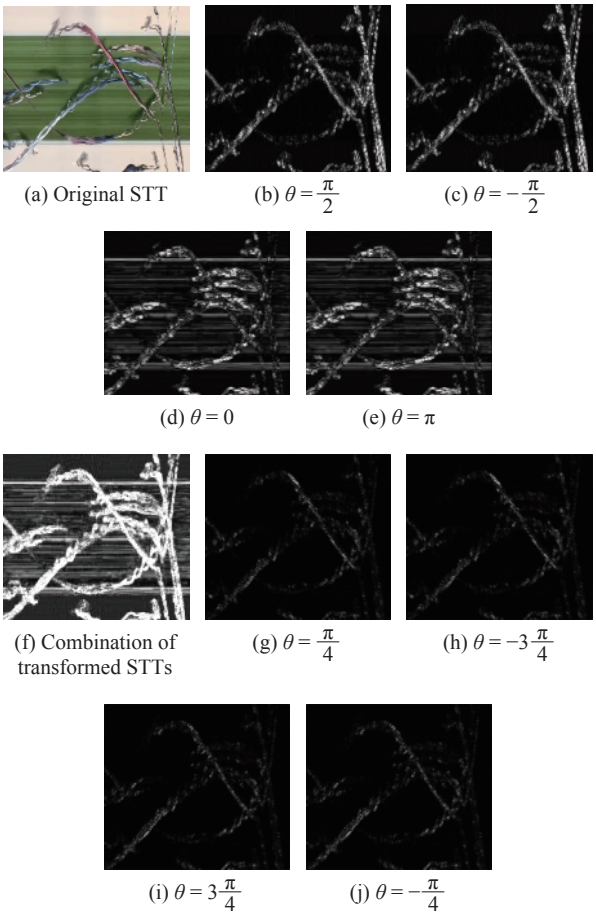


Fig. 6 Gabor filtering results along eight directions

The proposed algorithm only has to collect several STTs at certain positions, the amount of pixels needing to be analyzed is then reduced to $(N+1)wt + (M+1)ht$, thus the overall computational complexity is $O(n^2)$. Also, because patterns of STTs with varied signature values exhibit different behavioral types, by carefully selecting, some patterns could be modeled into a feature signature which could be used for further texture classification. Unlike the change detection algorithm introduced in the previous chapter, the classification of textures is capable of potentially labeling different scenarios in input video streams.

3 GLCM signaturing for classification

In order to achieve automatic warning of hazardous crowd behaviors, a spatio-temporal volume (STV) signature modeling method is proposed to detect crowd abnormality recorded in CCTV streams using the texture extraction algorithm proposed in Section 2. Once the optimal STTs are extracted, the gray level co-occurrence matrix (GLCM) can be formulated to measure the crowd behaviors identified. In this section, the proposed STT signatures based on the GLCM indices have been defined. The proposed model has shown a promising accuracy and efficiency in detecting crowd abnormal behaviors. It has been proven that the STT signatures are suitable descriptors for detecting certain crowd events, which provides an encouraging direction for real-time surveillance applications.

3.1 STT feature categorization

Depending on different construction patterns, STT features can be roughly classified into statistical texture features, model type texture features and signal domain texture features according to Junior et al.[33] Statistical texture features are obtained by transforming the gray scale values between a target pixel and its neighbors in the first-order, second-order and even higher-order filtering process to denote information – often described in the conventional terms of contrast, variance, etc. The most frequently used statistical texture features is the grey level co-occurrence matrix (GLCM)[34], which will be discussed in the next section. The model type texture features assume that a texture can be described by certain parameters controlled by probabilistic distribution models. How to recover the most accurate parameter values is the core issue of this approach. Benezeth et al.[35] proposed an algorithm using a hidden Markov model (HMM) associated with a spatio-temporal neighborhood co-occurrence matrix to describe the texture feature. In the signal domain texture features, textures are defined in a transformational domain by certain filters such as the wavelet[36]. It is based on the assumption that the energy distribution within the frequency domain can be used to classify textures.

The grey level co-occurrence matrix (GLCM), known as grey tone spatial dependency matrix, is first proposed by Haralick et al.[34] By definition, the GLCM is a statistic tabulation of the probability of different pixel grey scale values occurred in an image. In brief, assuming the gray scale of current image is divided into three levels, GLCM will store all the neighboring pairs of these three levels.

In this research, the GLCM patterns have been explored to test their performance on STT signature identification. The main strategy of this approach is to extract raw GLCM texture features from relevant STTs. Once these features are acquired, a signature could be modeled for classification purpose. A five-stage process flow of this approach is shown in Fig. 7.

In order to obtain the GLCM indices from a STT, the very first step is to transform a STT from RGB image to gray scale, and then the raw GLCM, labeled as G, can be calculated based on the algorithm introduced in [37]. In most cases, the gray scale value distribution of STTs is irregular, thus the obtained results of G are often asymmetric. According to the GLCM definition, G represents the gray-scale pair relations along one direction, the transposed matrix is then calculated to represent the rela-

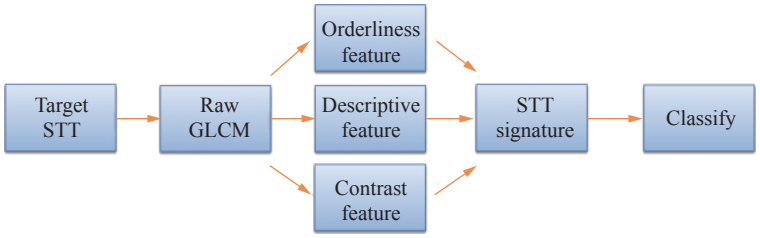


Fig. 7 Structure of the proposed approach

tion matrix along the opposite direction, and then the symmetric matrix S can be obtained by adding G' and G, to represent the complete relations along this direction. The next step is the normalization, where the probability matrix P is obtained from S by using (4).

$$P_{i,j} = \frac{S_{i,j}}{\sum_{i,j=0}^{N-1} S_{i,j}} \tag{4}$$

where the obtained i and j are the row and column indices of matrix S and P. Obtained probability matrix P has two properties: 1) According to the definition of GLCM algorithm, assuming that the gray scale value of the original image is divided into N levels, then the column and row numbers are also N. Thus, the more levels the gray scales are divided into, the larger N will be, which means the size of the GLCM will be larger. Also, the range of N is usually from 3 to 10. If it is too large, the GLCM will be sparse and its descriptive ability will be affected. In order to reduce the computation time and to avoid overly sparse GLCMs, a proper value of N should be selected. In this research, the value of N is set to 8 based on experiments. 2) P is symmetric along the diagonal. The diagonal elements represent pixels which do not have gray level differences, and the farther away from the diagonal, the greater the differences between the pixel gray levels. According to this property, patterns like the contrast can be readily retrieved in a look-up table style.

Next, texture patterns can be calculated from the probability matrix P. The resulting low level texture patterns are named here as contrast patterns, orderliness patterns, and descriptive statistical patterns.

3.2 Contrast patterns of GLCM

Contrast patterns describe how the gray scale value of current image varies in terms of contrast, dissimilarity, homogeneity and similarity. The farther the pixel pairs from the central diagonal line in P, the bigger the difference it represents within the gray scale, thus the contrast can be obtained by (5).

$$CON = \sum_{i,j=0}^{N-1} P_{i,j}(i-j)^2. \tag{5}$$

Similar to contrast, dissimilarity also represents difference in gray scale values, except it increases linearly instead of exponentially. Dissimilarity can be obtained by (6).

$$DIS = \sum_{i,j=0}^{N-1} P_{i,j}|i-j|. \tag{6}$$

Homogeneity is also called inverse different moment (IDM). On the contrary, homogeneity represents how consistent the contrast is, when the contrast of an image is low, the value of its homogeneity will be large. Equation (7) shows how to calculate homogeneity.

$$HOM = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}. \tag{7}$$

Similar to dissimilarity, the linear version of homogeneity can be obtained by (8).

$$SIM = \sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+|i-j|}. \tag{8}$$

Table 2 gives a comparison of the contrast related patterns for sample images. The GLCM window size is set to 50 by 50 pixels, where the direction is set to horizontal with the step size fixed at 1 pixel. The gray scale level number is set to 8. The patch in Table 2(a) is less contrastive than the one in Table 2(d), thus the result shows that patch in Table 2(a) has less GLCM contrast and dissimilarity values, and larger homogeneity and similarity values.

3.3 Orderliness patterns of GLCM

Orderliness related patterns describe how orderly or regular the distribution of gray scale values in an image is, including angular second moment, energy and entropy. The concept of angular second moment (ASM) comes from physics[21] for measuring rotational acceleration. ASM could be obtained using (9). Its value increases while the orderliness distribution is high.

$$ASM = \sum_{i,j=0}^{N-1} {P_{i,j}}^2. \tag{9}$$

Table 2 Comparison between texture patterns of STT patches

	Patch (a)	Patch (b)	Patch (c)	Patch (d)	Patch (e)	Patch (f)
						
Contrast	0.2437	0.3237	0.2669	0.6853	0.5735	0.6473
Dissimilarity	0.1922	0.2110	0.1935	0.3947	0.3645	0.4278
Homogeneity	0.9085	0.9049	0.9103	0.8302	0.8379	0.8078
Similarity	0.9101	0.9094	0.9132	0.8405	0.8459	0.8174
Angular second moment	0.3538	0.2134	0.4124	0.1853	0.2062	0.1767
Energy	0.5948	0.4619	0.6422	0.4304	0.4541	0.4203
Entropy	1.2977	2.0858	1.5294	2.3325	2.1747	2.3599
Mean	2.4933	4.7598	2.7865	4.0635	2.6410	2.8265
Variance	0.3859	3.7115	0.6728	2.5150	1.0509	2.8265
Standard deviation	0.6212	1.9265	0.8202	1.5859	1.0251	1.0729
Correlation	0.6843	0.9564	0.8016	0.8638	0.7272	0.7188
Normal	Yes	Yes	Yes	No	No	No

The energy equals to the square root of ASM, as (10). It is often used in fingerprint recognition[38] and plant classification[39].

$$ENR = \sqrt{ASM_{i,j}}. \tag{10}$$

On contrary to energy, entropy describes how irregular current gray scale distribution is, where the value of entropy decreases when the distribution is less orderly. Entropy can be expressed as (11).

$$ENT = \sum_{i,j=0}^{N-1} P_{i,j}(-\ln P_{i,j}). \tag{11}$$

In Table 2, the orderliness of six different images are measured. The patch in Table 2(a) clearly shows more regular patterns than the patch in Table 2(d), so it can be expected that the Entropy of the patch in Table 2(a) is less than the one in Table 2(d).

3.4 Descriptive statistical patterns of GLCM

Descriptive statistical related patterns consist of statistics derived from a GLCM matrix, including mean, variance and correlation. It needs to be emphasized that these patterns describe the statistical pixel pair relations, but not typical gray scale value explicitly. Two GLCM mean values can be obtained by using (12), note that because the probability matrix P is symmetric, the two mean values are identical.

$$\mu_i = \sum_{i,j=0}^{N-1} i(P_{i,j})$$
$$\mu_j = \sum_{i,j=0}^{N-1} j(P_{i,j}). \tag{12}$$

GLCM variance σ^2 and standard deviation σ can be obtained through (13).

$${\sigma_i}^2 = \sum_{i,j=0}^{N-1} P_{i,j}(i-\mu_i)^2$$
$${\sigma_j}^2 = \sum_{i,j=0}^{N-1} P_{i,j}(j-\mu_j)^2. \tag{13}$$

Finally, according to the calculated mean and variance, the GLCM correlation can be obtained by (14).

$$COR = \sum_{i,j=0}^{N-1} P_{i,j}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{({\sigma_i}^2)({\sigma_j}^2)}}\right]. \tag{14}$$

3.5 GLCM signature modeling

In this section, patterns of GLCM matrices are modeled as signatures for crowd motion classification. Six STT patches are extracted at different parts of the STV model in Table 2, Patches (a)–(c) are obtained from texture with normal motion, and Patches (d)–(f) are obtained from texture with abnormal motion. By comparing pattern values of normal and abnormal patches, the following patterns can be identified. Firstly, a texture patch at a normal state usually has lower contrast, en-

tropy and variance, e.g., Patches (a)–(c) all have lower contrast than Patches (d)–(f). Secondly, a texture patch with normal behavior usually has higher ASM value than patches at abnormal state. Thirdly, among all other patterns, contrast, ASM, entropy and variance show most significant changes between normal and abnormal states. Thus, these four GLCM-based patterns are selected as the most appropriate measures for detecting abnormal crowd states, and are denoted accordingly in Table 2.

Fig. 8(a) displays the gray scale image transformed from a STT obtained in Fig. 2(d), the actual test video is chosen from the University of Minnesota (UMN) dataset. All videos from this dataset start with a normal crowd scene followed by an abnormal event, mostly panic behavior. The ground truth of normal and abnormal behaviors is manually marked on Fig. 8(a), by using a color bars at the bottom of the figure. The grey color indicates normal state and the black color indicates abnormal state. It can be observed that different visual patterns of this figure match the labeled ground truth. It is expected that the differences of patterns will reflect the defined STT signatures too. According to the definition of STT, the column index represents the frame index in the original video, thus by summing up each column calculated by GLCM texture features, the change of GLCM feature patterns over time can be quantified and evaluated.

Figs. 8 (b)–8 (e) show the trends of contrast patterns of

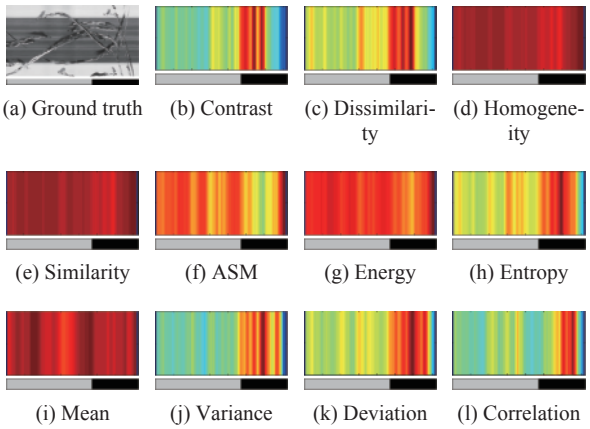


Fig. 8 Trends of GLCM patterns along time

the STT in Fig. 8 (a). As the anomaly occurs, patterns which describe pixel pair dissimilarity, such as contrast and dissimilarity, increase rapidly. However, patterns describing pixel pair similarity such as Homogeneity and Similarity do not change significantly. Figs. 8 (f)–8 (h) show the trends of orderliness patterns of Fig. 8(a). When the anomaly occurs, patterns describing image irregularity such as entropy increase quickly, while the angular second moment shows a significant drop though the energy holds steady. Figs. 8 (i)–8 (l) show the trends of statistic related pattern measures. When the anomaly occurs, the mean value does not show significant change. Variance, standard deviation and correlation value also change slightly. Hence, in summary, the contrast (CON), angular second moment (ASM), entropy (ENT) and variance (VAR) are selected as candidates for forming the STT signature vector for classification due to their salient variance magnitude. As the linear version of contrast, dissimilarity is discarded to control the dimension of the signature, the same decision process has been applied to the standard deviation and correlation. The final signature (SIG) for classification is modeled as (15).

$$SIG = [CON, ASM, ENT, VAR]. \tag{15}$$

4 Test and evaluations

In this section, an experimental system equipped with the devised signature model and process pipeline has been constructed to classify the crowd motion videos as shown in Fig. 9. The extracted STT is firstly filtered with the six-orientation Gabor transform to amplify the motion details. Once a STT is processed, it is divided into a collection of texture patches, and the patterns are extracted from these patches to model the signature for classification. In the classification phase, the texture patches are classified with a trained classifier using the modeled signatures. TAMURA texture patterns[40] are also utilized to model a signature for performance comparison. The values of coarseness, contrast, line likeness and regularity are modeled as a four dimensional TAMURA signature.

Several classifiers are implemented on these two patterns to assess the performance, including the K nearest neighbor (KNN), Naïve Bayes, discriminant analysis classifier (DAC), random forest and support vector machine (SVM). In the training stage for the classifier, extracted STTs with congestion and panic scenarios are divided into manually labeled texture patches to train the classifier. The texture patches for training are categorized into four different types, which are empty, normal, congested and panic. The empty texture contains no pedestrians but only background. The normal texture contains pedestrians walking casually in a scene. The congested texture contains pedestrians with slow moving velocity and high density. The panic texture contains pedestrians escaping of high velocity.

Once the classifier is trained, STTs for testing are firstly divided into patches and the patterns are extracted to model the signature for classification. The details of parameter setting for classifiers are as follows. The size of patches is set to 50 by 50 pixels. For the KNN, the number of neighbors is set to 4, since in training phase only four types of anomaly are defined. For the random forest classifier, the number of trees is set to 5. The parameters of Naïve Bays, DAC and SVM are set as default. One of classification results is shown as Fig. 10, the KNN classifier is applied on the GLCM signature. The blue line grid marks the boundary of each divided patch. The Dash

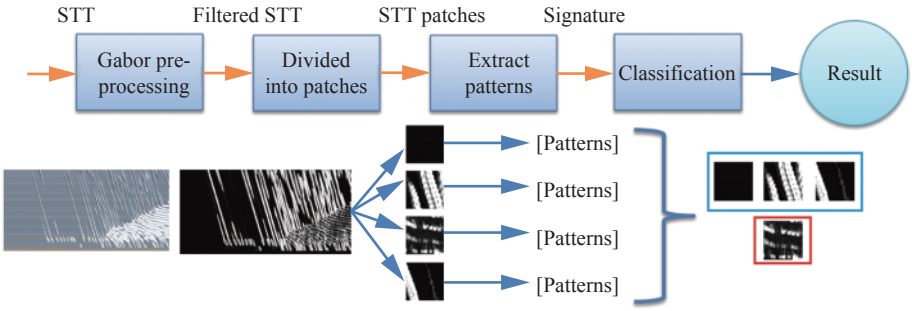


Fig. 9 Structure of proposed classification approach

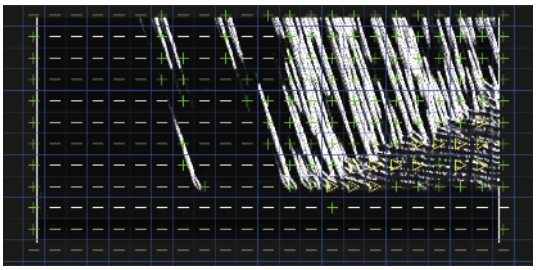


Fig. 10 Detection result using GLCM signature and KNN

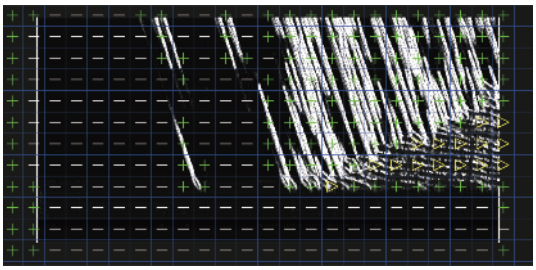
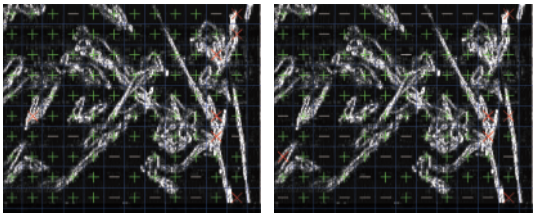


Fig. 11 Detection result using TAMURA signature and KNN



(a) GLCM and KNN (b) TAMURA and KNN

Fig. 12 Comparison of detection results on panic state

line stands for the empty texture, cross for the normal texture, triangle for the congested texture, and oblique cross for the panic texture. Since agent's velocity is higher in uncongested state, the spatial shifting along time will be larger than when it is in a congested state. As consequence, the texture stripe in STT will have larger slope value. On the contrary, textures containing congested behaviors will have parallel stripes with gently sloping value. Therefore, in visual expression, texture patch with more horizontal stripes stands for the congestion behavior, and the one with more vertical stripes stands for the normal states. In summary, congested texture patches have relatively smaller contrast, entropy, variance values and larger angular second moment value.

The TAMURA signatures for the same STT set have also been applied to the KNN classifier to compare the performance. The result is shown in Fig. 11. Comparing to Fig. 10, a number of texture patches with no motion patterns are marked as normal, and some with normal pedestrian behaviors are marked as congested, as highlighted in Fig. 11. The comparison indicated that the GLCM-based signature (feature vectors) outperformed the TAMURA in detecting crowd motion patterns.

The detection of panic scenes is also carried out. In Fig. 12, STTs extracted from the UMN dataset are processed using the proposed procedure. A comparison is made between the GLCM and TAMURA texture patterns. Fig. 12(a) shows the detection result using GLCM, and Fig. 12(b) shows the detection result using TAMURA. Similar to Figs. 10 and 11, agents with panic behavior are likely to have higher moving speed. Thus, the texture patch with panic behavior will show stripes in higher slope value.

In order to measure the performance, all sample test patches are manually labeled with the four texture types in the training phase. If the results equal to the labeled ground truths, then it is considered a correct detection, and the label value $C_{i,j}$ is set to 1, otherwise a failed one and the label value is set to 0. The detection accuracy A can be calculated using (16). Table 3 shows the accuracy between various combination of signatures and classifiers.

$$A = \frac{\sum_{i,j=0}^{N} C_{i,j}}{i \times j}. \tag{16}$$

5 Conclusions and future work

Real-time and effective monitoring of high density crowds for public safety is of increasing demand in the real world. In this research, a novel crowd anomaly detection framework is proposed that satisfies continuous feed-in of spatio-temporal information from live CCTVs. Novel STT selection, filtering, and feature modelling techniques have been devised and tested. Evaluation against state-of-the-art benchmarking systems yields satisfactory

Table 3 Accuracy of multiple signatures and classifiers combination

	Congestion 1	Congestion 2	Panic 1	Panic 2
GLCM+KNN	71.52%	79.59%	81.42%	63.33%
TAMURA+KNN	78.12%	87.75%	67.14%	63.33%
GLCM+SVM	58.68%	63.94%	68.57%	71.66%
TAMURA+SVM	82.98%	87.07%	68.57%	71.66%
GLCM+Naïve Bayes	81.94%	70.74%	67.14%	43.33%
TAMURA+Naïve Bayes	85.76%	85.03%	75.71%	60.83%
GLCM+DAC	80.55%	72.78%	78.57%	54.16%
TAMURA+DAC	82.98%	82.31%	74.28%	67.50%
GLCM+Random forest	74.30%	78.91%	68.57%	62.50%
TAMURA+Random forest	87.84%	88.43%	70.00%	70.00%

results with promising potential in further improving system adaptability under different application scenarios. High level semantic studies of the identified motion features will also be investigated in the future.

Acknowledgement

This research is funded by Chinese National Natural Science Foundation (No. 61671377) and Shaanxi Smart City Technology Project of Xianyang (No. 2017k01-25-5).

Open Access

References

[1] B. Solmaz, B. E. Moore, M. Shah. Identifying behaviors in crowd scenes using stability analysis for dynamical systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2064–2070, 2012. DOI: 10.1109/TPAMI.2012.123.

[2] B. Krausz, C. Bauckhage. Loveparade 2010: Automatic video analysis of a crowd disaster. *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 307–319, 2012. DOI: 10.1016/j.cviu.2011.08.006.

[3] X. Y. Cui, Q. S. Liu, M. C. Gao, D. N. Metaxas. Abnormal detection using interaction energy potentials. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, pp. 3161–3167, 2011. DOI: 10.1109/CVPR.2011.5995558.

[4] D. D. Ma, Q. Wang, Y. Yuan. Anomaly detection in crowd scene via online learning. In *Proceedings of the 14th International Conference on Internet Multimedia Computing and Service*, ACM, Xiamen, China, pp. 158–162, 2014. DOI: 10.1145/2632856.2632862.

[5] R. Raghavendra, A. Del Bue, M. Cristani, V. Murino. Optimizing interaction force for global anomaly detection in crowded scenes. In *Proceedings of IEEE International Conference on Computer Vision Workshops*, Barcelona, Spain, pp. 136–143, 2011. DOI: 10.1109/ICCVW.2011.6130235.

[6] A. Zeroual, N. Messai, S. Kechida, F. Hamdi. A piecewise switched linear approach for traffic flow modeling. *International Journal of Automation and Computing*, vol. 14, no. 6, pp. 729–741, 2017. DOI: 10.1007/s11633-017-1060-4.

[7] Z. C. Song, Y. Z. Ge, H. Duan, X. G. Qiu. Agent-based simulation systems for emergency management. *International Journal of Automation and Computing*, vol. 13, no. 2, pp. 89–98, 2016. DOI: 10.1007/s11633-016-0958-6.

[8] G. Csurka, C. R. Dance, L. X. Fan, J. Willamowski, C. Bray. Visual categorization with bags of keypoints. In *Proceedings of ECCV International Workshop on Statistical Learning in Computer Vision*, Grenoble, France, pp. 145–146, 2004.

[9] D. Helbing, P. Molnar. Social force model for pedestrian dynamics. *Physical Review E*, vol. 51, no. 5, pp. 4282–4294, 1995. DOI: 10.1103/PhysRevE.51.4282.

[10] W. Yan, Z. Zou, J. B. Xie, T. Liu, P. Q. Li. The detecting of abnormal crowd activities based on motion vector. *Optik*, vol. 166, pp. 248–256, 2018. DOI: 10.1016/j.ijleo.2017.11.187.

[11] Y. Hao, Y. Liu, J. L. Fan. A crowd behavior feature descriptor based on optical flow field. *Journal of Xi'an University of Posts and Telecommunications*, vol. 21, no. 6, pp. 55–59, 2016. DOI: 10.13682/j.issn.2095-6533.2016.06.011. (In Chinese)

[12] S. A. Niyogi, E. H. Adelson. Analyzing and recognizing walking figures in XYT. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 469–474, 1994. DOI: 10.1109/CVPR.1994.323868.

[13] M. Kass, A. Witkin, D. Terzopoulon. Snakes: Active contour models. *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988. DOI: 10.1007/BF00133570.

[14] J. Wang, Z. J. Xu. STV-based video feature processing for action recognition. *Signal Processing*, vol. 93, no. 8, pp. 2151–2168, 2012. DOI: 10.1016/j.sigpro.2012.06.009.

[15] C. Van Gemeren, R. Poppe, R. C. Veltkamp. Hands-on: deformable pose and motion models for spatiotemporal localization of fine-grained dyadic interactions. *EURASIP*

Journal on Image and Video Processing, vol. 2018, Article number 16, 2018. DOI: 10.1186/s13640-018-0255-0.

[16] X. F. Ji, Q. Q. Wu, Z. J. Ju, Y. Y. Wang. Study of human action recognition based on improved spatio-temporal features. *International Journal of Automation and Computing*, vol. 11, no. 5, pp. 500–509, 2014. DOI: 10.1007/s11633-014-0831-4.

[17] E. H. Aldelson, J. R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, vol. 2, no. 2, pp. 284–299, 1985. DOI: 10.1364/JOSAA.2.000284.

[18] Y. Iwashita, M. Petrou. Person identification from spatio-temporal volumes. In *Proceedings of the 23rd International Conference Image and Vision Computing*, IEEE, Christchurch, New Zealand, 2008. DOI: 10.1109/IVCNZ.2008.4762086.

[19] R. C. Bolles, H. H. Baker, D. H. Marimont. Epipolar-plane image analysis: an approach to determining structure from motion. *International Journal of Computer Vision*, vol. 1, no. 1, pp. 7–55, 1987. DOI: 10.1007/BF00128525.

[20] H. H. Baker, R. C. Bolles. Generalizing epipolar-plane image analysis on the spatiotemporal surface. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Ann Arbor, USA, pp. 33–49, 1988. DOI: 10.1109/CVPR.1988.196209.

[21] G. Kuhne G, S. Richter, M. Beier. Motion-based segmentation and contour-based classification of video objects. In *Proceedings of the 9th ACM international conference on Multimedia*, Ottawa, Canada, pp. 41–50, 2001. DOI: 10.1145/500141.500150.

[22] C. W. Ngo, T. C. Pong, R. T. Chin. Detection of gradual transitions through temporal slice analysis. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Fort Collins, USA, pp. 41–46, 1999. DOI: 10.1109/CVPR.1999.786914.

[23] Y. Hao, Z. J. Xu, J. Wang, Y. Liu, J. L. Fan. An approach to detect crowd panic behavior using flow-based feature. In *Proceedings of the 22nd International Conference on Automation and Computing*, IEEE, Colchester, UK, pp. 462–466, 2016. DOI: 10.1109/IConAC.2016.7604963.

[24] J. H. Xiang, H. Fan, J. Xu. Abnormal behavior detection based on spatial-temporal features. In *Proceedings of International Conference on Machine Learning and Cybernetics*, IEEE, Tianjin, China, pp. 871–876, 2013. DOI: 10.1109/ICMLC.2013.6890406.

[25] H. H. Alqaysi, S. Sasi. Detection of abnormal behavior in dynamic crowded gatherings. In *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop*, IEEE, Washington DC, USA, pp. 1–6, 2013. DOI: 10.1109/AIPR.2013.6749309.

[26] C. Li, Z. J. Han, Q. X. Ye, J. B. Jiao. Abnormal behavior detection via sparse reconstruction analysis of trajectory. In *Proceedings of the 6th International Conference on Image and Graphics*, IEEE, Hefei, China, pp. 807–810, 2011. DOI: 10.1109/ICIG.2011.104.

[27] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x.

[28] K. He, S. X. Wang. Study on denoising of fractal signal based on Shannon entropy. In *Proceedings of International Conference on Neural Networks and Signal Processing*, IEEE, Nanjing, China, pp. 751–755, 2003. DOI: 10.1109/ICNNSP.2003.1279384.

[29] S. Liang, Y. Ma, Y. Y. Huang, J. Guo, C. F. Jia. The scheme of detecting encoded malicious web pages based on information entropy. In *Proceedings of the 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, IEEE, Fukuoka, Japan, pp. 310–312, 2016. DOI: 10.1109/IMIS.2016.82.

[30] Z. J. Zhang, X. N. Wang, L. Sun. Mobile payment anomaly detection mechanism based on information entropy. *IET Networks*, vol. 5, no. 1, pp. 1–7, 2014. DOI: 10.1049/iet-net.2014.0101.

[31] D. K. Panda, S. Meher. Hierarchical background subtraction algorithm using Gabor filter. In *Proceedings of IEEE International Conference on Electronics, Computing and Communication Technologies*, Bangalore, India, pp. 1–6, 2015. DOI: 10.1109/CONECCT.2015.7383876.

[32] D. X. Zhou, H. Zhang, N. Ray. Texture based background subtraction. In *Proceedings of IEEE International Conference on Information and Automation*, Changsha, China, pp. 20–23, 2008. DOI: 10.1109/ICINFA.2008.4608070.

[33] J. C. S. J. Junior, S. R. Musse, C. R. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 66–77, 2010. DOI: 10.1109/MSP.2010.937394.

[34] R. M. Haralick, K. Shanmugam, I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973. DOI: 10.1109/TSMC.1973.4309314.

[35] Y. Benezeth, P. M. Jodoin, V. Saligrama, C. Rosenberger. Abnormal events detection based on spatio-temporal Co-occurences. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, pp. 2458–2465, 2009. DOI: 10.1109/CVPR.2009.5206686.

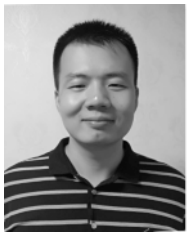
[36] J. B. Shen, X. G. Jin, C. Zhou, H. L. Zhao. Dynamic textures using wavelet analysis. In *Proceedings of International Conference*, Springer, Berlin Heidelberg, Germany, pp. 1070–1073, 2006. DOI: 10.1007/11736639_132.

[37] The GLCM Tutorial. [Online], Available: http://www.fp.ucalgary.ca/mhallbey/tutorial.htm, June 28, 2018.

[38] S. B. Nikam, S. Agarwal. Wavelet energy signature and GLCM features-based fingerprint anti-spoofing. In *Proceedings of the International Conference on Wavelet Analysis and Pattern Recognition*, IEEE, Hong Kong, China, pp. 717–723, 2008. DOI: 10.1109/ICWAPR.2008.4635872.

[39] G. Mukherjee, A. Chatterjee, B. Tudu. Study on the potential of combined GLCM features towards medicinal plant classification. In *Proceedings of the 2nd International Conference on Control, Instrumentation, Energy and Communication*, IEEE, Kolkata, India, pp. 98–102, 2016. DOI: 10.1109/CIEC.2016.7513746.

[40] R. K. Ranjan, A. Agrawal. Video summary based on F-sift, Tamura textural and middle level semantic feature. *Procedia Computer Science*, vol. 89, pp. 870–876, 2016. DOI: 10.1016/j.procs.2016.06.075.



Yu Hao received the B. Sc. degree in electronic engineering from Xidian University, China in 2008, and the M. Sc. degree in computer science from the Wichita State University, USA in 2011, and he is the Ph. D. degree candidate in computing and engineering from the University of Huddersfield, UK since 2015. Currently, he is a lecturer in School of Computer Science and

Technology, Xi′an University of Posts and Telecommunications, China. He has published about 7 refereed journal and conference papers during his Ph. D. program.

His research interest is crowd abnormal behavior analysis.

E-mail: haoyu@xupt.edu.cn

ORCID iD: 0000-0002-6820-5243

Zhi-Jie Xu received the B. Sc. degree in communication engineering from the Xi′an University of Science and Technology, China in 1991. After graduation, he has worked for one of the major Chinese Electrical and Machinery Companies – HH Yellow River Ltd for four years as an electronics engineer. He then traveled to the UK and spent a year working in a robotics labratory in Derby, UK. In 1996, he registered and became a postgraduate student at the University of Derby, UK. His research topic is virtual reality for manufacturing simulations. In 2000, he has completed his Ph. D. study and immediately been offered a tenured academic post at the University of Huddersfield, UK. He has published over 100 peer-reviewed journal and conference papers as well as editing 5 books in the relevant fields. He has supervised 11 postgraduate (including 8 Ph. D.) students to completion and been continuously winning substantial research and development grants in his career to date. He is a member of the IEEE, Institution of Engineering and Technology (IET), British Computer Society (BCS), The British Machine Vision Association (BMVA) and a fellow of Higher Education Academy (HEA). In addition, he has been serving as an editor, reviewer and chair for many prestigious academic journals and conferences.

His research interests include visual computing, vision systems, data science and machine learning.

E-mail: z.xu@hud.ac.uk (Corresponding author)

ORCID iD: 0000-0002-0524-5926

Ying Liu received the Ph. D. degree in computer vision from the Monash University, Australia in 2007. And she worked as a post doctor researcher at Nanyang Technological University, Singapore until 2010. She is the chief engineer of Shaanxi Forensic Science Digital Information Laboratory Research Center, China since 2012. Currently, she is the assistant dean of School of Communications and Information Engineering at Xi′an University of Posts and Telecommunications, China. She has published over 60 peer-reviewed journal and conference papers in the relevant fields. She was grant annual best paper of Pattern Recognition and Tier A paper from Australia Research Council.

Her research interest include pattern recognition, machine learning and forensic science.

E-mail: ly_ yolanda@sina.com

ORCID iD: 0000-0003-1796-8045

Jing Wang received the B. Sc. degree in machine and electronic technology from the Xidian University, China in 2006. After graduation, he was appointed as software engineer and carried out development work on computer vision (CV)-based quality control systems, such as assembly line monitoring and industrial robotic controls. In 2008, he began his postgraduate study at the University of Huddersfield and received his Ph. D. degree in computer vision from University of Huddersfield, UK in 2012. He then became a research fellow and carried out independent researches on image processing, analysing and understanding. Since 2008, He has published more than 20 journal and conference papers in the relative fields. He is a member of the British Machine Vision Association (BMVA) and British Computer Society (BCS). He has also served as chair and editor for *the International Conference on Automation and Computing*.

His research interest is real-world applications of computer vision systems.

E-mail: jing.wang@shu.ac.uk

ORCID iD: 0000-0002-8579-8765

Jiu-Lun Fan received the B. Sc. and M. Sc. degrees in mathematics from the Shaanxi Normal University, China in 1985 and 1988, respectively, and the Ph. D. degree in electronic engineering from the Xidian University, China in 1998. Currently, he is the president of Xi′an University of Posts and Telecommunications, China since 2015. He has published over 200 peer-reviewed journal and conference papers in the relevant fields.

His research interests include signal processing, pattern recognition and communications security.

E-mail: jiulunf@xupt.edu.cn